Navy Personnel Research and Development Center
San Diego, California 92152-7250 TN-94-2 October 1993

A Procedure for Linear Polychotomous Scoring of Test Items

J. Bradford Sympson

NPRDC-TN-94-2 October 1993

A Procedure for Linear Polychotomous Scoring of Test Items

J. Bradford Sympson

Reviewed by
Daniel O. Segall

Approved and released by
W. A. Sands
Director, Personnel Systems Departmenı

Navy Personnel Research and Development Center
San Diego, CA 92152-7250

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
October 1993

3. REPORT TYPE AND DATE COVERED
Final—October 1988-September 1991

4. TITLE AND SUBTITLE
A Procedure for Linear Polychotomous Scoring of Test Items

5. FUNDING NUMBERS
Program Element: 0601153N
Work Unit: R4204

6. AUTHOR(S)
J. Bradford Sympson

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Navy Personnel Research and Development Center
San Diego, California 92152-7250

8. PERFORMING ORGANIZATION REPORT NUMBER
NPRDC-TN-94-2

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of the Assistant Secretary of Defense (FM&P)
The Pentagon
Washington, DC 20301-3210

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

Functional Area:	Personnel
Product Line:	Computerized Testing
Effort:	Computerized Adaptive Testing

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE
A

13. ABSTRACT *(Maximum 200 words)*

A new procedure for scoring aptitude and achievement tests is described. The procedure involves the computation of scoring weights that are then associated with the response categories of test items. When tests are scored using these scoring weights, test reliability increases.

The new procedure is called *polyweighting*. The scoring weights are called *polyweights*. Polyweights are computed using an iterative algorithm that is described in this report.

In addition to describing the numerical procedure for computing polyweights, the report shows, and discusses in detail, output from the computer program POLY. The example demonstrates how polyweighting can be used to calibrate and score test items drawn from an item bank that is too large to allow each examinee to answer every question. It also demonstrates how polyweighting increases the internal consistency reliability of aptitude and achievement tests.

14. SUBJECT TERMS
Selection, classification, training, testing, item scoring, polychotomous scoring, polytomous scoring

15. NUMBER OF PAGES
27

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT
UNCLASSIFIED

18. SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

19. SECURITY CLASSIFICATION OF ABSTRACT
UNCLASSIFIED

20. LIMITATION OF ABSTRACT
UNLIMITED

Foreword

This technical note describes a new approach to scoring multiple-choice items. The approach appears promising as it does nct require any assumptions about "latent" abilities, item-set dimensionality, or the mathematical form of the regression of item responses on unobservable variables. Moreover, it allows the development of large item banks in situations where only a portion of the items can be administered to each examinee.

Results reported in this technical note were originally presented at the Office of Naval Research Contractors' Meeting on Model-based Psychological Measurement, which was held in Iowa City, IA, in May of 1988. It is being published at this time for archival purposes.

The research described here was conducted under the Navy Personnel Research and Development Center Independent Research and Independent Exploratory Development (IR/IED) Programs. Additional funding was provided by the Joint Service Computerized Adaptive Testing-Armed Services Vocational Aptitude Battery (CAT-ASVAB) Program, which is sponsored by the Office of the Assistant Secretary of Defense (FM&P). Preparation of this document was funded by the Office of Naval Research (Code 1142) under the Navy Laboratory Participation Program (Program Element 0601153N, Work Order R4204).

W. A. SANDS
Director, Personnel Systems Department

Summary

Problem

Conventional methods for scoring aptitude and achievement tests that are used in selecting, classifying, and training military personnel discard useful information about an examinee's ability/ skill level. Information is lost whenever the original responses to test questions are classified only as "right" or "wrong." Additional information can be obtained by considering the difficulty level of the questions answered correctly and by taking into account which particular wrong answers were selected.

Objective

The objective of this effort was to develop new procedures for scoring aptitude and achievement tests that will increase the reliability and validity of those tests.

Approach

A new approach to scoring multiple-choice items was developed. The procedure is not based on Item Response Theory (IRT), and does not require any assumptions regarding "latent" abilities, the dimensionality of the set of items analyzed, or the mathematical form of the regression of item responses on unobservable variables. The procedure does assume that the individuals included in an item analysis are randomly sampled from the examinee population of interest. The procedure is characterized as "linear" because each examinee's score is a linear function of category scoring weights and category-response indicators.

The new scoring procedure is called *polyweighting*. In this procedure, the scoring weights obtained for an item are independent of the difficulty of other items included in the item analysis and the weights are bounded so that examinees who give the correct answer to an item will always receive the most credit. For each correct answer, and each wrong answer selected by 100 or more examinees, the category scoring weight is approximately equal to the mean percentile rank among examinees selecting the category. For each wrong answer selected by fewer than 100 examinees, the scoring weight for the category is "regressed" toward the mean percentile rank among examinees who chose any wrong answer on the item.

Results

A detailed example of an item analysis using the computer program "POLY" is presented, demonstrating several features of the scoring procedure. In particular, the example shows that polyweighting does not require a fully-crossed data matrix (one in which all examinees have been administered all questions) and that polyweighting increases coefficient-α for the set of items analyzed.

Conclusions

The scoring procedure described in this report provides an improved foundation for scoring aptitude and achievement tests. It makes few assumptions about the available data and can be

implemented with smaller sample sizes than are required for IRT scoring. Users of the procedure can elect either to keep tests at their current length and increase score reliability, or to reduce test length to save testing time while maintaining reliabilities at current levels.

Recommendations

Organizations that administer aptitude and/or achievement tests for purposes of personnel selection, classification, or training should consider whether this new scoring procedure can be usefully applied to their tests.

Contents

List of Figures

Introduction

Polychotomous scoring of test items, while not widely practiced, has a lengthy history. Haladyna and Sympson (1988) reviewed that history and distinguished between two approaches to polychotomous scoring. One approach involves the assignment of differential scoring weights to item response categories. In this approach to polychotomous scoring, the test score is a linear function of the examinee's item response vector.

One method of linear polychotomous scoring has the unique property of maximizing coefficient-α (Cronbach, 1951) for the set of items calibrated (Guttman, 1941; Lord, 1958). This scoring procedure has been referred to by various names, including *reciprocal averages scaling* (Horst, 1935), *optimal scaling* (Bock, 1960), and *dual scaling* (Nishisato, 1980). Since these names are not suggestive of the method's primary distinguishing characteristic, the present author refers to it as "max-alpha" (MA) scaling.

MA scaling has two drawbacks as an approach to polychotomous item scoring. First, the scoring weights that are derived for an item depend on the difficulty level of the other items that are calibrated at the same time. If an item is calibrated along with a set of easy items, the obtained scoring weights will be different than if the item were calibrated along with a set of difficult items. Second, in order to maximize α, the MA method often assigns weights to wrong answers that exceed the weight assigned to the correct response.

The second approach to polychotomous item scoring discussed by Sympson and Haladyna (1988) has a shorter history. This approach derives from Item Response Theory (IRT). IRT models for polychotomous calibration of multiple-choice items have been introduced within the past two decades (Bock, 1972; Samejima, 1979; Sympson, 1981, 1983, 1993; Thissen & Steinberg, 1984). In this approach to polychotomous item scoring, the test score is a nonlinear function of the examinee's item response vector.

If the set of items calibrated with a polychotomous IRT model is unidimensional, and if the chosen model fits the items, the model parameters for any one item will be independent of the parameters obtained for other items. However, if the assumed model is not correct, IRT item parameters are dependent on both the examinee population that is sampled and the set of items that is calibrated. IRT calibration methods require fairly large samples ($N \geq 1000$ per item) in order to provide stable results.

This report introduces a new approach to linear polychotomous scoring of test items. The approach is similar to MA scaling in some regards, but provides scoring weights for a given item that are independent of the difficulty of other items in the analysis. Moreover, the scoring weights are bounded so that an examinee can never receive more credit for an incorrect response than for a correct response.

Approach

Computing Polyweights

Polyweighting is a scoring procedure that uses a different scoring weight for each item response category. An examinee's *polyscore* is equal to the mean of the scoring weights of the categories chosen by the examinee. Polyweighting does not require the assumptions of IRT, and can be applied with smaller samples than are commonly required with IRT models. Polyweighting does require that item calibration be carried out with a random sample of examinees from the population of interest.

Unlike some scoring methods, polyweighting gives the examinee more credit for correct answers to difficult questions and less credit for correct answers to easy questions. Conversely, polyweighting penalizes the examinee more heavily for wrong answers to easy questions than for wrong answers to difficult questions. This may be contrasted with number/proportion-correct (PC) scoring and with scoring under the 1-parameter and 2-parameter logistic IRT models. The latter scoring methods assign scores to examinees in a manner that renders the scores independent of the difficulty of the questions answered correctly or incorrectly (Birnbaum, 1968, p. 458).

In polyweighting, the scoring weights assigned to item-response categories are referred to as *polyweights*. An iterative procedure must be used to derive polyweights for a set of items. The procedure is as follows:

1. Each examinee in the calibration sample is assigned a provisional score equal to the examinee's proportion correct among items the examinee was administered. It is assumed that different examinees may have been administered different items during data collection, but that an adequate number of examinees (e.g., 100 or more) was administered each "set" of items. It is also assumed that item-sets were assigned to examinees randomly, so that each "item-set group" is randomly equivalent to other examinee groups.

2. Since PC scores for examinees who are administered different item-sets are not directly comparable (due to variation in difficulties and other characteristics of the items administered), each examinee's PC score is converted to a percentile rank relative to those examinees who were administered the same item-set. This is equivalent to an equipercentile equating of PC scores from different item-sets (Angoff, 1971, p. 563). For each examinee, his/her percentile rank is the proportion of examinees, among those who were administered the same item-set, who obtained a PC score that was less than or equal to the PC score obtained by the given examinee, multiplied by 100.

3. For each item, the mean percentile rank among examinees who chose each possible response category is determined. This computation includes all examinees who were administered a given item, even if they were administered different item-sets. At this point, if the mean percentile rank among examinees who chose the correct answer for a given item is less than the mean percentile rank among all examinees who were administered the item, the item is deleted from the analysis. This is equivalent to deleting an item if the *point-biserial correlation* (Henrysson, 1971, p. 142) between the correct answer and examinee percentile ranks becomes negative.

4. For all items and all response categories, provisional polyweights are computed as follows:

a. For each correct answer, the provisional polyweight is equal to the mean percentile rank among examinees choosing the category, rounded to the nearest integer.

b. For each wrong answer chosen by 100 or more examinees, the provisional polyweight is equal to the mean percentile rank among examinees choosing the category, rounded to the nearest integer.

c. For each wrong answer chosen by fewer than 100 examinees, the provisional polyweight is a rounded linear combination of the mean percentile rank among examinees choosing the category and the mean percentile rank among examinees choosing any wrong answer on the item. For these categories, the polyweight for category *j* of item *i* is equal to

$$W_{ij} = \bar{R}_{i(w)} + \left[\frac{N_{ij}}{100} \right]^{1/2} (\bar{R}_{ij} - \bar{R}_{i(w)}) \quad , \qquad (1)$$

rounded to the nearest integer. In Equation 1, $\bar{R}_{i(w)}$ is the mean percentile rank among examinees choosing any wrong answer on item *i*, $\bar{R}_{ij}$ is the mean-percentile rank among examinees choosing category *j*, and N_{ij} is the number of examinees choosing category *j*.

5. Since examinee percentile ranks range from a minimum possible value of $100(1/N)$ to a maximum possible value of 100, the provisional polyweights can assume any integer value from 0 to 100. For a given item, if the provisional polyweight for an incorrect response is found to equal or exceed the provisional polyweight for the correct response, the polyweight for the incorrect response is set equal to 1 less than the polyweight for the correct response. Thus, under polyweighting, an examinee can never receive more credit for an incorrect answer than for a correct answer.

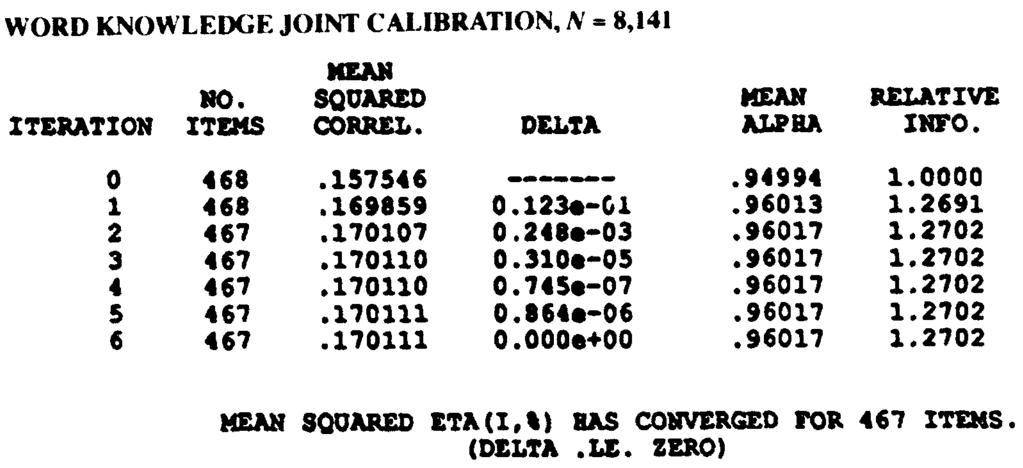
6. Given the provisional polyweights for all response categories, provisional examinee polyscores are computed. As stated earlier, an examinee's polyscore is equal to the mean of the polyweights of the categories chosen by the examinee. Since polyscores, like all raw test scores, are not comparable between examinees who have taken different item-sets, the provisional polyscores are converted to percentile ranks within each group of examinees who have been administered the same set of items.

7. Given the new percentile ranks for all examinees, the iterative procedure returns to Step 3, above. Steps 3 through 6 are repeated until the mean squared correlation ratio between items and percentile ranks stops increasing.

Example and Discussion

Output From the Computer Program POLY

Figures 1 through 3 show selected portions of a "Primary Output" file generated by the computer program POLY. In this example, polyweights were derived for 467 items that had been administered to 8,141 applicants for military service. Each applicant was administered one of three 86-item vocabulary tests and one of six 35-item vocabulary tests. Thus, there were 18 different item-sets administered to the examinees, with 121 items in each item-set.



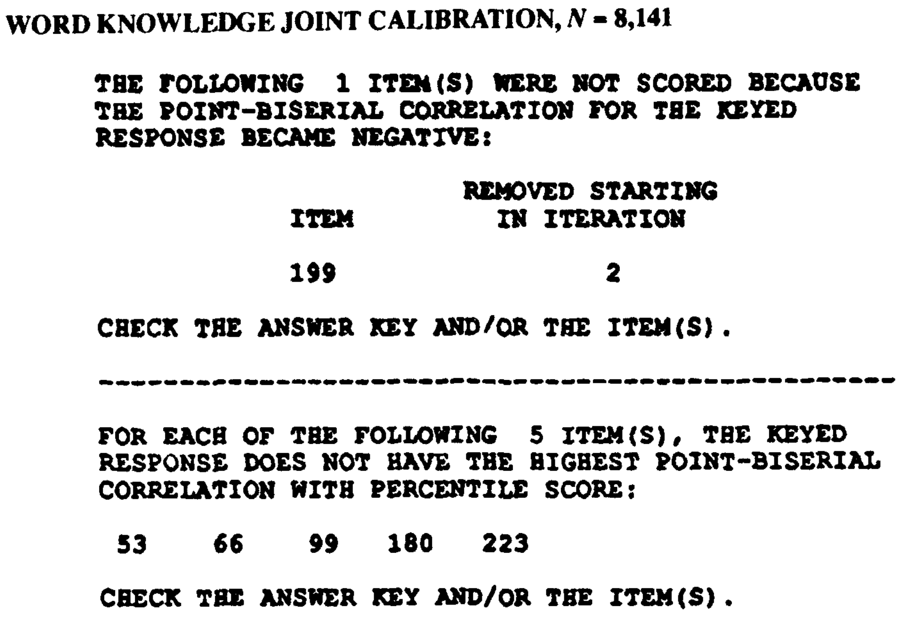
WORD KNOWLEDGE JOINT CALIBRATION, *N* = 8,141

```
                       MEAN
                NO.    SQUARED                           MEAN      RELATIVE
ITERATION      ITEMS   CORREL.        DELTA              ALPHA      INFO.

    0           468    .157546       -------            .94994     1.0000
    1           468    .169859     0.123e-01            .96013     1.2691
    2           467    .170107     0.248e-03            .96017     1.2702
    3           467    .170110     0.310e-05            .96017     1.2702
    4           467    .170110     0.745e-07            .96017     1.2702
    5           467    .170111     0.864e-06            .96017     1.2702
    6           467    .170111     0.000e+00            .96017     1.2702

              MEAN SQUARED ETA(I,%) HAS CONVERGED FOR 467 ITEMS.
                             (DELTA .LE. ZERO)
```

Figure 1. Convergence data.



WORD KNOWLEDGE JOINT CALIBRATION, *N* = 8,141

```
THE FOLLOWING  1 ITEM(S) WERE NOT SCORED BECAUSE
THE POINT-BISERIAL CORRELATION FOR THE KEYED
RESPONSE BECAME NEGATIVE:

                        REMOVED STARTING
          ITEM            IN ITERATION

          199                  2

CHECK THE ANSWER KEY AND/OR THE ITEM(S).

---------------------------------------------------

FOR EACH OF THE FOLLOWING  5 ITEM(S), THE KEYED
RESPONSE DOES NOT HAVE THE HIGHEST POINT-BISERIAL
CORRELATION WITH PERCENTILE SCORE:

 53    66    99   180   223

CHECK THE ANSWER KEY AND/OR THE ITEM(S).
```

Figure 2. Diagnostic information from the Primary Output File generated by POLY.

WORD KNOWLEDGE JOINT CALIBRATION, *N* = 8,141

```
                              ITEM  65

                     (WK, BOOK 1, ITEM 65       )

                     ----------------------------
                       SUMMARY ITEM ANALYSIS
                     (** INDICATES KEYED RESPONSE)
                     ----------------------------

                           SCORING   %-ILE    %-ILE    ADJ.
CAT.      FREQ.    PROP.   WEIGHT    MEAN     S.D.     PROP.      R(C,%)
 0        5312   0.6525   ------    50.15    28.86    0.         ------
 1           9   0.0011     6.00     2.29     1.46    0.0032    -0.0936
 2          16   0.0020     9.00     9.81    17.74    0.0057    -0.1053
 3 **     2779   0.3414    51.00    50.90    28.48    0.9823     0.1963
 4           8   0.0010     9.00    12.13    25.07    0.0028    -0.0701
 5           7   0.0009     9.00    11.55    16.03    0.0025    -0.0666
 6           3   0.0004     8.00    10.39    12.80    0.0011    -0.0449
 7           7   0.0009     6.00     1.13     1.22    0.0025    -0.0845
1-7       2829   0.3475   ------    50.14    28.88    1.0000     ------

                       ETA(I,%) = 0.1972
```

WORD KNOWLEDGE JOINT CALIBRATION, *N* = 8,141

```
                              ITEM  66

                     (WK, BOOK 1, ITEM 66       )

                     ----------------------------
                       SUMMARY ITEM ANALYSIS
                     (** INDICATES KEYED RESPONSE)
                     ----------------------------

                           SCORING   %-ILE    %-ILE    ADJ.
CAT.      FREQ.    PROP.   WEIGHT    MEAN     S.D.     PROP.      R(C,%)
 0        5312   0.6525   ------    50.15    28.86    0.         ------
 1         334   0.0410    39.00    39.27    25.95    0.1181    -0.1377
 2         108   0.0133    25.00    24.80    21.05    0.0382    -0.1748
 3         179   0.0220    29.00    29.12    25.17    0.0633    -0.1891
 4 **      402   0.0494    61.00    61.20    35.78    0.1421     0.1559
 5        1787   0.2195    54.00    53.61    25.60    0.6317     0.1576
 6          11   0.0014    44.00    36.54    27.21    0.0039    -0.0294
 7           8   0.0010    35.00     2.61     4.09    0.0028    -0.0876
1-7       2829   0.3475   ------    50.14    28.88    1.0000     ------

                       ETA(I,%) = 0.3437
```

Figure 3. Two examples of the summary item analysis.

Figure 1 shows "Convergence Data" from the example Primary Output file. Column 1 in Figure 1 gives iteration numbers. Iteration 0 is the iteration in which each examinee is assigned a raw score equal to his/her proportion correct among the items the examinee was administered. Subsequent iterations use provisional polyweights to compute examinee scores.

Column 2 in Figure 1 indicates how many items were included in the analysis during each iteration. In this example, one item was deleted, starting in iteration 2, because the point-biserial correlation between the item's correct response and percentile rank scores became negative.

Column 3 in Figure 1 gives the mean, over all retained items, of the squared correlation ratio between an item and percentile rank scores. In iteration 0, the value reported is the mean squared point-biserial correlation between correct responses and percentile ranks. For a given item, the squared point-biserial correlation for the correct response is equal to the proportion of variance in percentile ranks that is accounted for by knowing whether each examinee has selected the correct response. The point-biserial correlation is a widely-used index of item discriminating power when item scoring is dichotomous.

In subsequent iterations, the value reported in column 3 is the mean squared η coefficient between an item and percentile ranks (Lord & Novick, 1968, p. 263). The squared η coefficient between an item and percentile rank scores indicates the proportion of variance in percentile ranks that is accounted for by knowing which particular response category an examinee has selected. The η coefficient for an item can never be smaller than the correct-answer point-biserial correlation. If there is any variation among the score means for the wrong-answer categories, the η coefficient for an item will be larger than the point-biserial correlation.

Column 4 in Figure 1 shows the change (δ) in the mean squared correlation ratio between iterations. This quantity serves as the convergence criterion in POLY runs. When δ becomes so small that it cannot be distinguished from zero, or if δ becomes negative, the iterations are terminated.

Column 5 in Figure 1 gives the mean value of coefficient-α in the analysis. Since there were 18 item-sets in this example, each value reported in column 5 is the mean of 18 α coefficients. As the example shows, polyweighting increased the mean value of α for the item-sets analyzed. The fact that scores based on polyweights have higher α coefficients than do number/proportion-correct scores implies that test scores based on polyweighting will correlate more highly with domain scores and will have higher alternate-form reliabilities.

Column 6 in Figure 1 gives an index of "relative information." This index is based on the Spearman-Brown formula (Lord & Novick, 1968, p. 112). The Spearman-Brown formula gives the reliability of a lengthened test as a function of the initial reliability of the test and the proportionate increase in test length that is anticipated. However, rather than use the Spearman-Brown formula to predict reliability, one can rearrange the formula and use it to determine how much a given test would have to be inc^r^eased in length in order to obtain a specified level of reliability (Nishisato, 1980, p. 118).

In POLY, the relative information index is set equal to 1.0000 in iteration 0. Subsequent to iteration 0, the formula used for computing relative information is

$$H = \frac{\bar{\alpha}_p \ (1 - \bar{\alpha}_d)}{\bar{\alpha}_d \ (1 - \bar{\alpha}_p)} \quad . \tag{2}$$

where $\bar{\alpha}_d$ is the mean value of coefficient-α obtained under PC scoring (iteration 0) and $\bar{\alpha}_p$ is the mean value of coefficient-α obtained under polyweighting. This information index indicates the proportionate increase in test length that would be required in order to achieve the same reliability under PC scoring that has been achieved using polyweighting.

In the example shown in Figure 1, the POLY run terminated after iteration 6. At that time, the mean value of coefficient-α was .96017. When this value is substituted for $\bar{\alpha}_p$ in Equation 2, and the initial mean α of .94994 is substituted for $\bar{\alpha}_d$, the obtained final value of the relative information index is 1.2702. This indicates that a typical item-set in this analysis would have to be increased in length by 27% (i.e., from 121 items to 154 items) in order to achieve the level of reliability under PC scoring that was achieved using polyweighting.

Figure 2 shows diagnostic information from the Primary Output file generated by POLY. In the example, Item 199 was deleted (not scored) starting in iteration 2, because the point-biserial correlation between the item's correct response and percentile rank scores was negative at the end of iteration 1. Items 53, 66, 99, 180, and 223 were scored, but have been flagged for special attention because each of these items had at least one incorrect answer with a positive point-biserial correlation that was larger than the point-biserial correlation for the correct answer.

Figure 3 shows two examples of the "Summary Item Analysis" provided for each item in the Primary Output file. Items 65 and 66 were selected as examples because Item 65 is quite easy and Item 66 is quite difficult. Moreover, Item 66 is one of the items that was flagged by POLY as needing special attention. Below the item-number for each item, a 25-character item-identification string is printed in parentheses. The user specifies this string for each item in the analysis. In Figure 3, both items came from Word Knowledge (WK) test-booklet number 1.

The columns headed "CAT." in Figure 3 contain response-category identification numbers. Category 0 is a pseudo-category that corresponds to "Not Administered." If an examinee's data-record contains a zero response-code for a given item, POLY does not use that item in computing the examinee's polyscore. In POLY, eight categories are available as scored categories. The user must indicate the number of categories that are present for each item in the analysis. The number of categories can vary from item to item. In the examples in Figure 3, Categories 1 through 5 correspond to choices "A" through "E" in these 5-alternative multiple-choice items.

For the items in Figure 3, Category 6 corresponds to "Omit." For each item, a response-code of 6 was entered in the examinee data-record if the examinee did not answer the item, but he/she answered at least one item that appeared later in the same test booklet. Category 7 corresponds to "Not Reached." A response-code of 7 was entered in an examinee's data-record if the examinee did not answer a given item, and he/she did not answer any subsequent items in the same test booklet. This use of Categories 6 and 7 is specific to our example. During a POLY run, the last two categories for an item are treated no differently than the other response categories (except Category 0).

In Figure 3, the last value that appears in the columns headed "CAT." identifies a composite pseudo-category that collapses all actual response categories into one. In the examples, this pseudo-category is labeled "1-7" because Items 65 and 66 were each specified to have seven categories. Summary statistics derived from all examinees who were administered a given item (i.e., from all categories other than Category 0) are associated with this pseudo-category.

The entries in the columns headed "FREQ." in Figure 3 indicate how many examinees were associated with each response category. The frequency shown for Category 0 is the number of examinees who were not administered the item (*N* = 5,312 for Items 65 and 66). The frequency shown for the composite pseudo-category (Category 1-7 in the example) is the number of examinees who were administered the item (*N* = 2,829 for Items 65 and 66). The other frequencies in this column correspond to the categories indicated in column 1. The double-asterisk (**) that appears between columns 1 and 2 identifies the keyed (correct) response for each item.

In Figure 3, the entries in the columns headed "PROP." indicate the proportion of the examinee sample that was associated with each response category. In these columns, the proportions are based on the entire examinee sample (*N* = 8,141 in the example). The entries in the columns headed "ADJ. PROP." (column 7) are adjusted proportions. These proportions are based on just the examinees who were actually administered an item. Thus, for Item 65, .3414 of the examinee sample gave the keyed response (Category 3, or "C"). However, since only .3475 of the sample was administered the item, the adjusted proportion for Category 3 is .9823, indicating that Item 65 was quite easy. This may be contrasted with the adjusted proportion for the correct answer to Item 66 (.1421), which indicates that Item 66 was quite difficult.

The columns headed "%-ILE MEAN" and "%-ILE S.D." in Figure 3 give the means and standard deviations of percentile rank scores among examinees associated with each response category. Means and standard deviations are computed for the pseudo-categories (Categories 0 and, in this example, 1-7) so that the user can check for obvious violations of the requirement that each item be administered to a random sample from the examinee population. For each item, the means and standard deviations for the two pseudo-categories should be similar. If they are not, it suggests that randomly equivalent item-set groups were not achieved.

In the example, the mean percentile rank among the 2,779 examinees selecting the correct response on Item 65 was 50.90. The mean percentile rank among individuals choosing a wrong answer on this item ranged from a high of 12.13 among the 8 individuals who chose Category 4 ("D") to a low of 1.13 among the 7 individuals who did not reach the item (Category 7).

The mean percentile rank among the 402 examinees who selected the correct response on Item 66 was 61.20. The mean percentile rank among individuals choosing a wrong answer on Item 66 ranged from a high of 53.61 among the 1787 individuals who chose Category 5 ("E"), to a low of 2.61 among the 8 individuals who did not reach the item (Category 7). Most of the category means for wrong-answers are substantially higher for Item 66 than for Item 65.

Final (iteration 6) polyweights for Items 65 and 66 are shown in the columns labeled "Scoring Weight" in Figure 3. For the keyed response categories, and for wrong answers selected by 100 or more examinees, these weights are the category means from iteration 5, rounded to the nearest integer. For wrong-answer categories selected by fewer than 100 examinees, the scoring weights were obtained by inserting percentile means from iteration 5 into Equation 1, and rounding the resulting W_{ij} values to the nearest integer.

For both example items, none of the wrong-answer scoring weights exceeded the polyweight for the item's keyed response, so none of the wrong-answer weights had to be bounded. If any weight had been bounded (set equal to 1 less than the polyweight for the keyed answer), a single asterisk (*) would have appeared to the right of the bounded weight.

Consideration of the scoring weights shown in Figure 3 gives an indication of the impact of polyweighting. An examinee who answers Item 66 correctly will receive more credit than an examinee who answers Item 65 correctly (61 vs. 51). Conversely, an examinee who answers Item 65 incorrectly will be penalized more heavily than an examinee who answers Item 66 incorrectly (a score of 9 or less if Item 65 is answered incorrectly vs. a score of at least 25 if Item 66 is answered incorrectly).

The columns headed "R(C,%)" in Figure 3 contain point-biserial correlations between individual response categories and percentile rank scores. In the case of Item 65, there is only one positive point-biserial, the one for the keyed answer. In the case of Item 66, there are two positive point-biserials, one associated with the keyed answer, and one associated with Category 5 ("E"). In fact, the point-biserial for Category 5 is slightly higher than the point-biserial for the keyed answer. This is why Item 66 was mentioned in the diagnostic output shown in Figure 2.

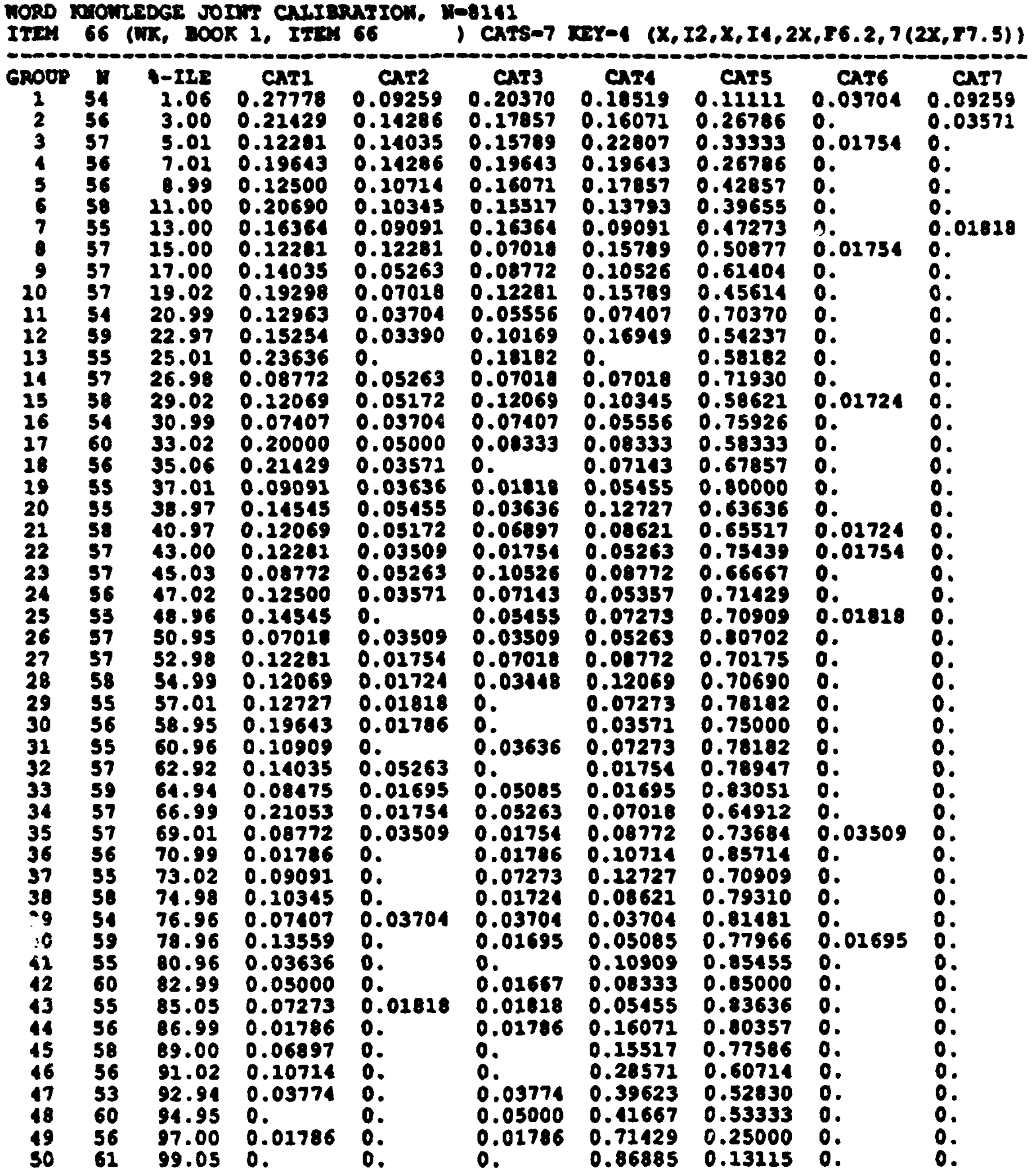
In Figure 3, the last summary statistic printed for each item is labeled "ETA(I,%)." This is the eta (η) coefficient for the item. For Item 65, the η coefficient is only slightly larger than the point-biserial for the keyed response (.1972 vs. .1963), which indicates that polychotomous scoring of this item will add very little to measurement precision. This is not unexpected, since Item 65 is very easy and few wrong answers are observed. For Item 66, the η coefficient is substantially larger than the point-biserial for the keyed response (.3437 vs. .1559), indicating that polychotomous scoring of this item will provide useful additional information about an individual's percentile rank within the examinee population.

An item such as Item 66 would often be discarded in a traditional item analysis because of the relatively large positive point-biserial correlation for Category 5, a wrong answer. However, by using polyweighting, this apparently bad item can be retained and used to gather useful information about examinee ability. As indicated by the percentile mean for Category 5, examinees who select this category are, on the average, of higher ability than examinees who select the other wrong answers. This fact is taken into account when the item is scored using the polyweights shown in Figure 3.

One would not want to use Item 66 in a test without further investigation of its psychometric properties and its content. To aid in this process, POLY can generate an "Endorsement-Rate Tables" file. The endorsement-rate table for Item 66 is shown in Figure 4. In order to create tables like the one shown in Figure 4, POLY can divide the examinee sample into as many as 100 ability groups, based on their percentile ranks. Then, using only the examinees who were administered a particular item, the proportion of examinees who gave each response is computed within each group. In the example, POLY was instructed to form 50 ability groups.

In Figure 4, the mean percentile rank within each ability group is shown in the column labeled "%-ILE." The columns labeled "CAT1" through "CAT7" contain the proportions selecting each category, within each ability group. It is notable that the proportion selecting the correct answer

(Category 4) is below chance level (i.e., below .20) over virtually all of the ability range from the 1st to the 89th percentiles. Only in the top decile of this examinee sample does the proportion of examinees that select the keyed answer start increasing, with the highest ability group (the top 2%) selecting the keyed answer about 87% of the time.



```
WORD KNOWLEDGE JOINT CALIBRATION, N=8141
ITEM   66 (WK, BOOK 1, ITEM 66       ) CATS=7 KEY=4 (X,I2,X,I4,2X,F6.2,7(2X,F7.5))
-----------------------------------------------------------------------------------
GROUP  N   %-ILE   CAT1     CAT2     CAT3     CAT4     CAT5     CAT6     CAT7
  1   54   1.06  0.27778  0.09259  0.20370  0.18519  0.11111  0.03704  0.09259
  2   56   3.00  0.21429  0.14286  0.17857  0.16071  0.26786  0.       0.03571
  3   57   5.01  0.12281  0.14035  0.15789  0.22807  0.33333  0.01754  0.
  4   56   7.01  0.19643  0.14286  0.19643  0.19643  0.26786  0.       0.
  5   56   8.99  0.12500  0.10714  0.16071  0.17857  0.42857  0.       0.
  6   58  11.00  0.20690  0.10345  0.15517  0.13793  0.39655  0.       0.
  7   55  13.00  0.16364  0.09091  0.16364  0.09091  0.47273  0.       0.01818
  8   57  15.00  0.12281  0.12281  0.07018  0.15789  0.50877  0.01754  0.
  9   57  17.00  0.14035  0.05263  0.08772  0.10526  0.61404  0.       0.
 10   57  19.02  0.19298  0.07018  0.12281  0.15789  0.45614  0.       0.
 11   54  20.99  0.12963  0.03704  0.05556  0.07407  0.70370  0.       0.
 12   59  22.97  0.15254  0.03390  0.10169  0.16949  0.54237  0.       0.
 13   55  25.01  0.23636  0.       0.18182  0.       0.58182  0.       0.
 14   57  26.98  0.08772  0.05263  0.07018  0.07018  0.71930  0.       0.
 15   58  29.02  0.12069  0.05172  0.12069  0.10345  0.58621  0.01724  0.
 16   54  30.99  0.07407  0.03704  0.07407  0.05556  0.75926  0.       0.
 17   60  33.02  0.20000  0.05000  0.08333  0.08333  0.58333  0.       0.
 18   56  35.06  0.21429  0.03571  0.       0.07143  0.67857  0.       0.
 19   55  37.01  0.09091  0.03636  0.01818  0.05455  0.80000  0.       0.
 20   55  38.97  0.14545  0.05455  0.03636  0.12727  0.63636  0.       0.
 21   58  40.97  0.12069  0.05172  0.06897  0.08621  0.65517  0.01724  0.
 22   57  43.00  0.12281  0.03509  0.01754  0.05263  0.75439  0.01754  0.
 23   57  45.03  0.08772  0.05263  0.10526  0.08772  0.66667  0.       0.
 24   56  47.02  0.12500  0.03571  0.07143  0.05357  0.71429  0.       0.
 25   55  48.96  0.14545  0.       0.05455  0.07273  0.70909  0.01818  0.
 26   57  50.95  0.07018  0.03509  0.03509  0.05263  0.80702  0.       0.
 27   57  52.98  0.12281  0.01754  0.07018  0.08772  0.70175  0.       0.
 28   58  54.99  0.12069  0.01724  0.03448  0.12069  0.70690  0.       0.
 29   55  57.01  0.12727  0.01818  0.       0.07273  0.78182  0.       0.
 30   56  58.95  0.19643  0.01786  0.       0.03571  0.75000  0.       0.
 31   55  60.96  0.10909  0.       0.03636  0.07273  0.78182  0.       0.
 32   57  62.92  0.14035  0.05263  0.       0.01754  0.78947  0.       0.
 33   59  64.94  0.08475  0.01695  0.05085  0.01695  0.83051  0.       0.
 34   57  66.99  0.21053  0.01754  0.05263  0.07018  0.64912  0.       0.
 35   57  69.01  0.08772  0.03509  0.01754  0.08772  0.73684  0.03509  0.
 36   56  70.99  0.01786  0.       0.01786  0.10714  0.85714  0.       0.
 37   55  73.02  0.09091  0.       0.07273  0.12727  0.70909  0.       0.
 38   58  74.98  0.10345  0.       0.01724  0.08621  0.79310  0.       0.
 39   54  76.96  0.07407  0.03704  0.03704  0.03704  0.81481  0.       0.
 40   59  78.96  0.13559  0.       0.01695  0.05085  0.77966  0.01695  0.
 41   55  80.96  0.03636  0.       0.       0.10909  0.85455  0.       0.
 42   60  82.99  0.05000  0.       0.01667  0.08333  0.85000  0.       0.
 43   55  85.05  0.07273  0.01818  0.01818  0.05455  0.83636  0.       0.
 44   56  86.99  0.01786  0.       0.01786  0.16071  0.80357  0.       0.
 45   58  89.00  0.06897  0.       0.       0.15517  0.77586  0.       0.
 46   56  91.02  0.10714  0.       0.       0.28571  0.60714  0.       0.
 47   53  92.94  0.03774  0.       0.03774  0.39623  0.52830  0.       0.
 48   60  94.95  0.       0.       0.05000  0.41667  0.53333  0.       0.
 49   56  97.00  0.01786  0.       0.01786  0.71429  0.25000  0.       0.
 50   61  99.05  0.       0.       0.       0.86885  0.13115  0.       0.
```

Figure 4. Endorsement-rate table for Item 66.

In Figure 4, the endorsement rates for wrong answers usually decline as ability level increases, though the rate of decline varies between response categories. A pronounced exception to this pattern is observed for Category 5, where the endorsement rate goes from chance level, among the lowest ability groups, to over .80 within ability groups near the 75th percentile. In the top decile of the examinee sample, the endorsement rate for Category 5 finally starts dropping, declining to about .13 in the highest ability group.

To aid interpretation of the endorsement-rate table shown in Figure 4, Figures 5 through 11 show graphic plots of the category endorsement rates for Item 66. The computer program POLY does not provide this type of plot, but such plots can be generated using the endorsement-rate tables that are available from POLY. The plots in Figures 5 through 11 also show fitted functions that smooth and interpolate the plotted endorsement rates. It is clear from Figures 10 and 11 that few examinees omit or do not reach Item 66 (Categories 6 and 7). Most examinees select Category 5 (Figure 9) and only the most able examinees select Category 4 (Figure 8) with greater than chance frequency.

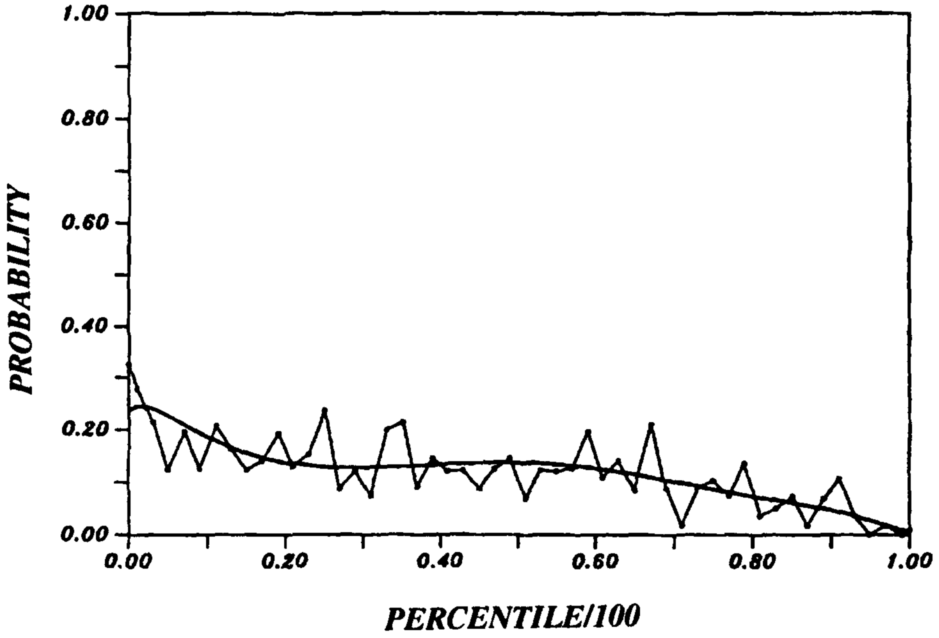


Figure 5. Plo‹ of Category 1 endorsement rates for Item 66.

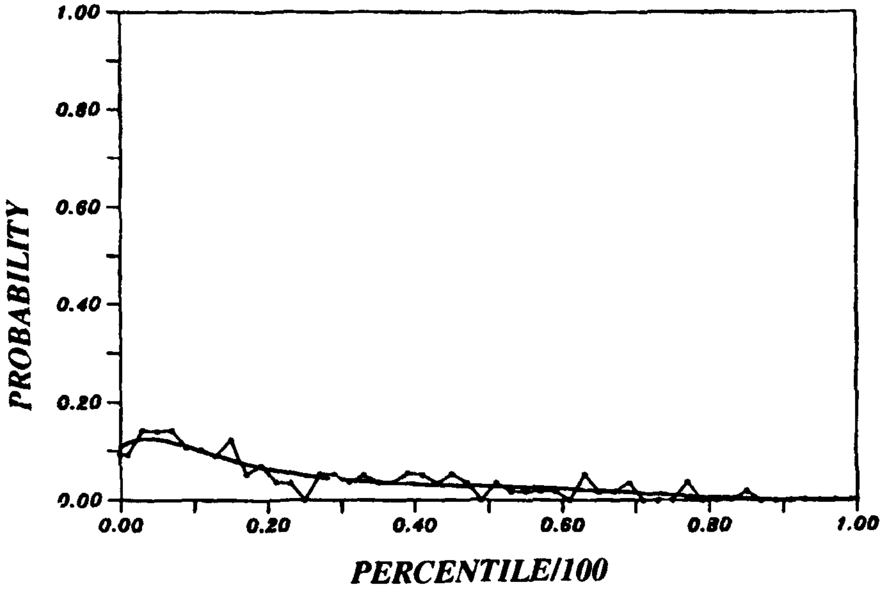


Figure 6. Plot of Category 2 endorsement rates for Item 66.

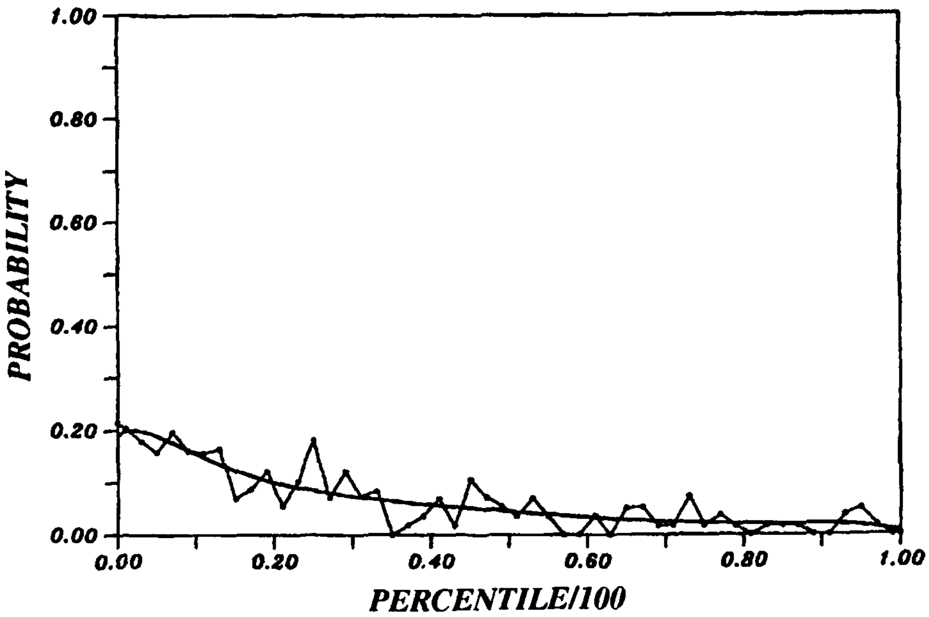


Figure 7. Plot of Category 3 endorsement rates for Item 66.

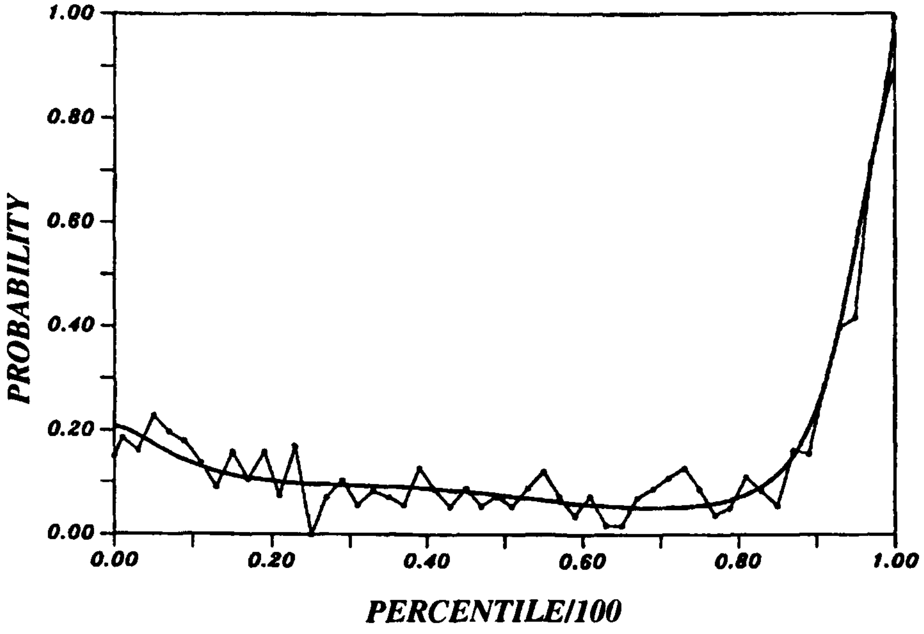


Figure 8. Plot of Category 4 endorsement rates for Item 66.

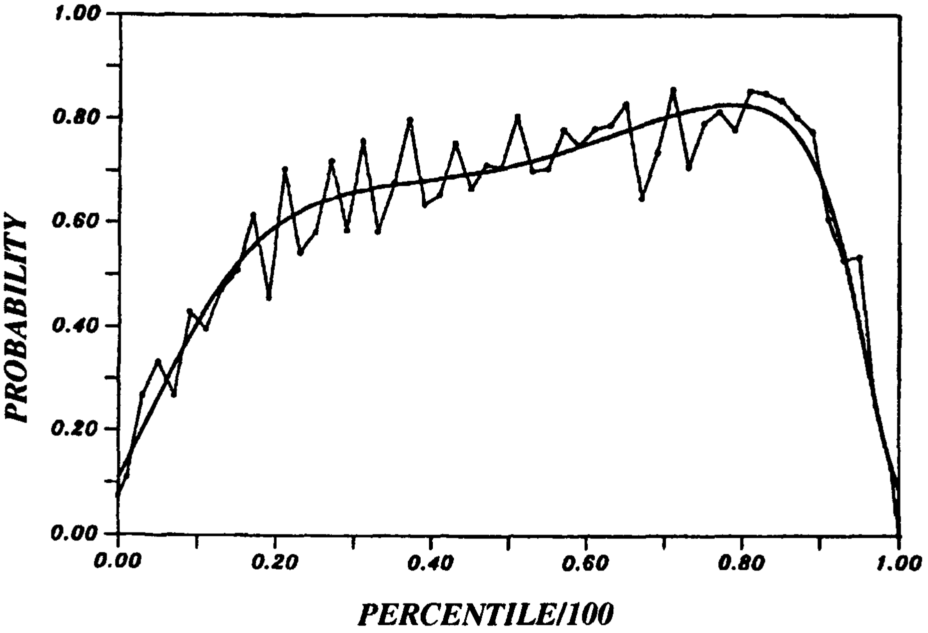


Figure 9. Plot of Category 5 endorsement rates for Item 66.

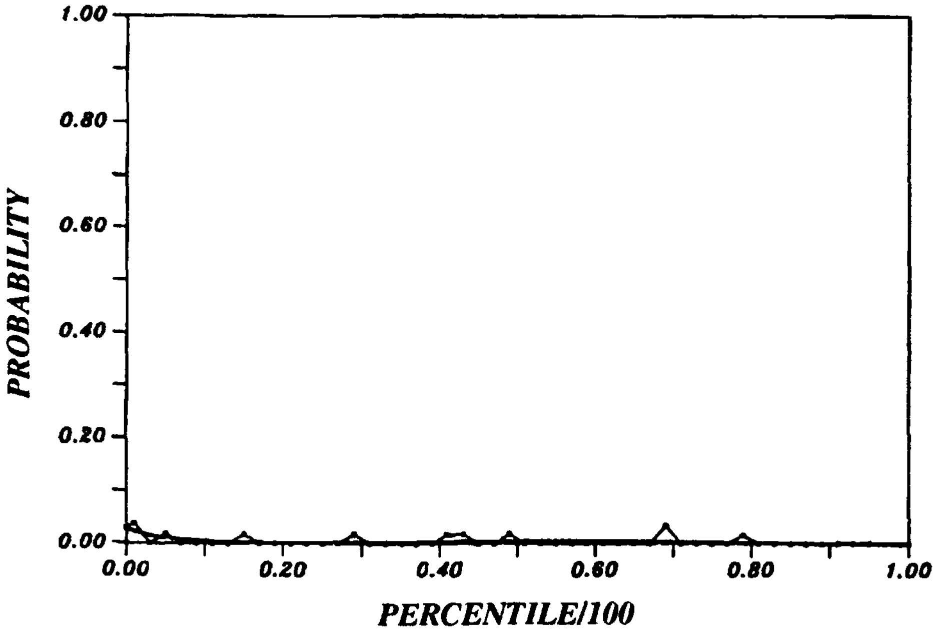


Figure 10. Plot of Category 6 endorsement rates for Item 66.

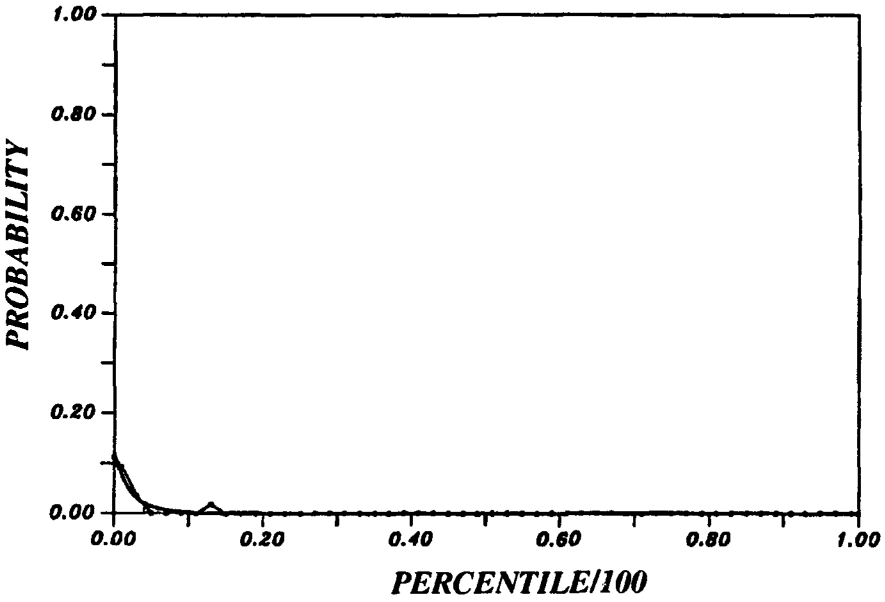


Figure 11. Plot of Category 7 endorsement rates for Item 66.

The most important step in the evaluation of an unusual item is a careful inspection of its content. Some information about Item 66 may be of interest. First, as mentioned previously, Item 66 is a 5-alternative multiple-choice WK item. This item asks the examinee to select the best synonym for "respite." The keyed answer is "rest" (Category 4). Category 5, the very popular wrong answer, is "grudge." It appears that all but the most knowledgeable examinees were fooled by the presence of the word "spite" within the item stem. Further inspection of Item 66 gave no indication of a problem with the content of the item, so its use in a polychotomously-scored test seems appropriate.

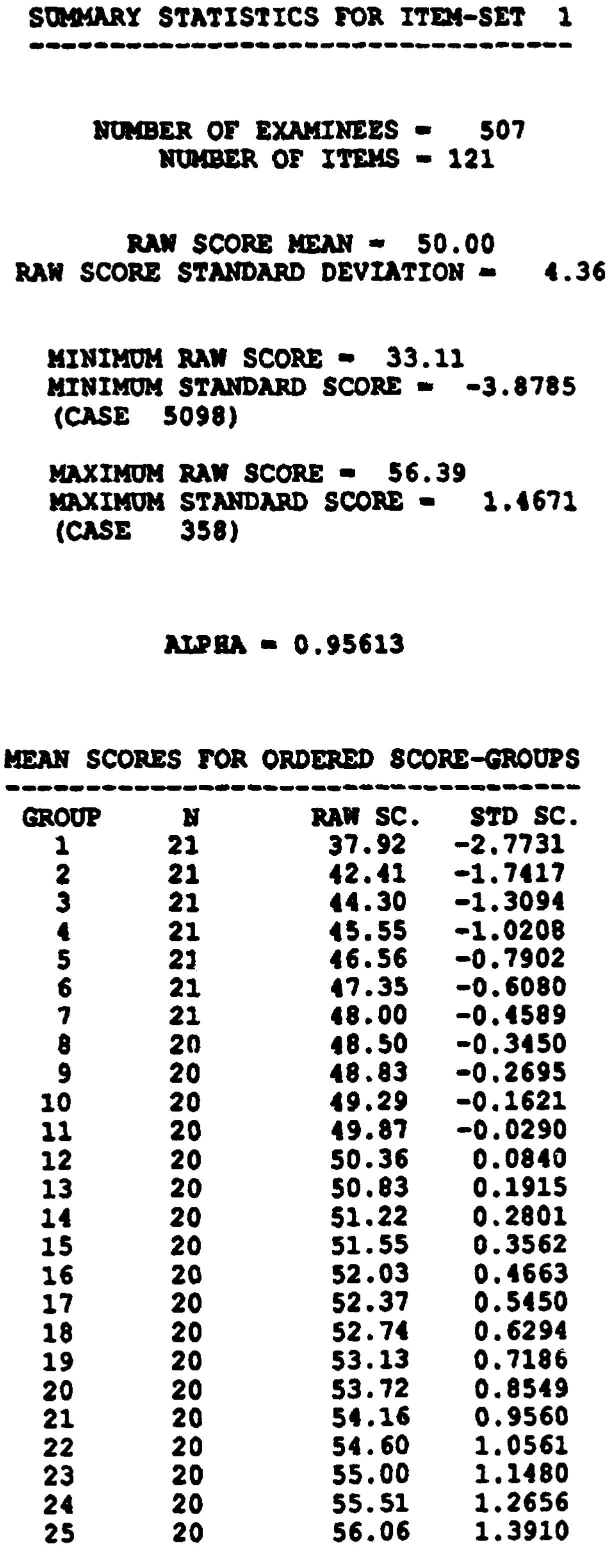
The Primary Output file generated by POLY also contains summary statistics for each item-set administered. An example is shown in Figure 12. As mentioned earlier, there were 18 item-sets administered in this item calibration example. Figure 12 shows summary statistics for one of these item-sets. Statistics provided by POLY include the number of examinees who were administered the item-set, the number of items in the item-set, the mean and standard deviation of raw polyscores for the item-set, the minimum raw/standardized polyscore observed, the maximum raw/standardized polyscore observed, and coefficient-α for the item-set. In addition to these summary statistics, a table giving the mean raw polyscore and the mean standardized polyscore for each of 25 equal-frequency score groups is printed. This table allows the user to gain an impression of the shape of the distribution of raw/standardized polyscores for each item-set. In Figure 12, it is clear that the distribution of raw/standardized polyscores was skewed left for item-set 1.

Conclusion

This concludes our discussion of example output from the program POLY. The example demonstrates that polyweighting can be used to calibrate large item-pools in which different examinees have been administered different test questions. Until now, it was necessary to adopt the assumptions of IRT in order to analyze this type of data. Such assumptions are no longer necessary.

The example also demonstrates that polyweighting increases the internal-consistency reliability of the item-sets (tests) to which it is applied. Available research (Sympson & Davison, 1993; Sympson & Haladyna, 1993) demonstrates that such reliability increases hold up well in new samples of examinees from the same population.

WORD KNOWLEDGE JOINT CALIBRATION, N = 8, 141



```
            SUMMARY STATISTICS FOR ITEM-SET  1
            ----------------------------------


                NUMBER OF EXAMINEES =   507
                    NUMBER OF ITEMS = 121


                  RAW SCORE MEAN =  50.00
           RAW SCORE STANDARD DEVIATION =   4.36


             MINIMUM RAW SCORE =  33.11
             MINIMUM STANDARD SCORE =  -3.8785
             (CASE  5098)

             MAXIMUM RAW SCORE =  56.39
             MAXIMUM STANDARD SCORE =   1.4671
             (CASE   358)


                    ALPHA = 0.95613


          MEAN SCORES FOR ORDERED SCORE-GROUPS
          ------------------------------------
           GROUP      N         RAW SC.    STD SC.
             1       21          37.92    -2.7731
             2       21          42.41    -1.7417
             3       21          44.30    -1.3094
             4       21          45.55    -1.0208
             5       21          46.56    -0.7902
             6       21          47.35    -0.6080
             7       21          48.00    -0.4589
             8       20          48.50    -0.3450
             9       20          48.83    -0.2695
            10       20          49.29    -0.1621
            11       20          49.87    -0.0290
            12       20          50.36     0.0840
            13       20          50.83     0.1915
            14       20          51.22     0.2801
            15       20          51.55     0.3562
            16       20          52.03     0.4663
            17       20          52.37     0.5450
            18       20          52.74     0.6294
            19       20          53.13     0.7186
            20       20          53.72     0.8549
            21       20          54.16     0.9560
            22       20          54.60     1.0561
            23       20          55.00     1.1480
            24       20          55.51     1.2656
            25       20          56.06     1.3910
```

Figure 12. Summary statistics for Item-set 1.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (chapters 17-20). Reading, MA: Addison-Wesley.

Bock, R. D. (1960). *Methods and applications of optimal scaling* (Research Memorandum 25). Chapel Hill, NC: Psychometric Laboratory, University of North Carolina.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Guttman, L. (1941). The quantification of a class of attributes: A theory and method of scale construction. In P. Horst (Ed.), *Prediction of personal adjustment* (Bulletin 48, pp. 321-348). New York: Social Science Research Council.

Haladyna, T. M., & Sympson, J. B. (1988, April). Empirically-based polychotomous scoring of multiple-choice test items: Historical overview. Paper presented in C. E. Davis (Chair), *New Developments in Polychotomous Item Scoring and Modeling*. Symposium conducted at the annual meeting of the American Educational Research Association, New Orleans.

Henrysson, S. (1971). Gathering, analyzing, and using data on test items. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 130-159). Washington, DC: American Council on Education.

Horst, P. (1935). Measuring complex attitudes. *Journal of Social Psychology, 16*, 369-374.

Lord, F. M. (1958). Some relations between Guttman's principal components of scale analysis and other psychometric theory. *Psychometrika, 23*, 291-296.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Nishisato, S. (1980). *Analysis of categorical data: Dual scaling and its applications*. Toronto, Canada: University of Toronto Press.

Samejima, F. (1979). *A new family of models for the multiple-choice item* (Research Report 79-4). Knoxville, TN: University of Tennessee, Department of Psychology.

Sympson, J. B. (1981, October). *A nominal model for IRT item calibration*. Talk given at the Office of Naval Research Conference on Model-based Psychological Measurement, Millington, TN.

Sympson, J. B. (1983, June). *A new item response theory model for calibrating multiple-choice items*. Paper presented at the meeting of the Psychometric Society, Los Angeles, CA.

Sympson, J. B. (1993). *Extracting information from wrong answers in computerized adaptive testing* (NPRDC-TN-94-1). San Diego: Navy Personnel Research and Development Center.

Sympson J. B.. & Davison, M. L. (1993). *Reducing test length with polychotomous scoring* (NPRDC-TN-94-4). San Diego: Navy Personnel Research and Development Center.

Sympson, J. B., & Haladyna, T. M. (1993). *An evaluation of "polyweighting" in domain-referenced testing* (NPRDC-TN-94-3). San Diego: Navy Personnel Research and Development Center.

Thissen, D., & Steinberg, L. (1984). A response model for multiple-choice items. *Psychometrika, 49*, 501-519.

Distribution List

Distribution:
Office of the Assistant Secretary of Defense (FM&P)
Office of Naval Research (Code 1142) (3)
Defense Technical Information Center (DTIC) (12)

Copy to:
Office of Naval Research (Code 20P), (Code 222), (Code 10)
Naval Training Systems Center, Technical Library (5)
Office of Naval Research, London
Director, Naval Reserve Officers Training Corps Division (Code N1)
Chief of Naval Education and Training (L01) (2)
Curriculum and Instructional Standards Office, Fleet Training Center, Norfolk, VA
Chief of Naval Operations (N71)
Director, Recruiting and Retention Programs Division (PERS-23)
Commanding Officer, Sea-Based Weapons and Advanced Tactics School, Pacific
Commanding Officer, Naval Health Sciences Education and Training Command, Bethesda, MD
Marine Corps Research, Development, and Acquisition Command (MCRDAC), Quantico, VA
AISTA (PERI II), ARI
Armstrong Laboratory, Human Resources Directorate (AL/HR), Brooks AFB, TX
Armstrong Laboratory, Human Resources Directorate (AL/HRMIM), Brooks AFB, TX
Armstrong Laboratory AL/HR-DOKL Technical Library, Brooks, AFB, TX
Library, Coast Guard Headquarters
Superintendent, Naval Post Graduate School
Director of Research, U.S. Naval Academy
Naval Education and Training Program (NETPMSA, Code 047), Pensacola (N. N. Perry)